日英報道記事からの訳語対応推定における複数の推定尺度の利用*

日野 浩平† 宇津呂 武仁†† 中川 聖一†

†豊橋技術科学大学 工学部 情報工学系 ††京都大学大学院 情報学研究科

1 はじめに

近年，ウェブ上の日本国内の新聞社などのサイトにおいては，日本語だけでなく英語で書かれた報道記事も掲載しており，これらの英語記事においては，同一時期の日本語記事とほぼ同じ内容の報道が含まれている．本研究では，これらの報道記事のページから日本語および英語など，異なった言語で書かれた文書を収集し，多種多様な分野について，分野固有の固有名詞 (固有表現) や事象・言い回しなどの翻訳知識を自動または半自動で獲得するというアプローチをとる [Utsuro03]．

これまでに研究されてきた翻訳知識獲得の手法は，大きく，対訳コーパスからの獲得手法とコンパラブルコーパスからの獲得手法に分けられる [Matsumoto00]．通常，対訳コーパスにおいては，文の対応の情報を利用することにより，片方の言語におけるタームや表現について，もう一方の言語における訳の候補が比較的少数に絞られるため，翻訳知識の獲得は相対的には容易といえる．ただし，そのような対訳コーパスを人手で整備する必要がある点が短所である．一方，コンパラブルコーパスからの獲得 (例えば，[Fung98]) では，各タームの周囲の文脈の類似性を言語横断して測定することにより，訳語対応の推定が行われる．情報源となるコーパスを用意するコストは小さくて済むが，翻訳知識の獲得は相対的に難しく，高性能に翻訳知識獲得を行うのは容易ではない．これらの従来のアプローチと比べると，報道記事を用いる本研究のアプローチは，情報源となるコーパスを用意するコストについては，コンパラブルコーパスを用いるアプローチと同等となる．しかし，同時期の報道記事を用いるため，片方の言語におけるタームや表現の訳がもう一方の言語の記事の方に出現する可能性が高く，翻訳知識の獲得が相対的に容易になるという大きな利点がある．

本研究における日英関連報道記事からの翻訳知識獲得の流れを図 1 に示す [Utsuro03]． まず，翻訳知識獲得のための情報源収集を目的として，同時期に日英二言語で書かれたウェブ上の新聞社やテレビ局のサイトから，報道内容がほぼ同一もしくは密接に関連した日本語記事および英語記事を検索する [浜本 03]．そして，取得された関連記事対に対し，内容的に対応する翻訳部分の推定を行い，その推定範囲から二言語間の訳語対応を推定し，訳語対の獲得を行う．ここで，訳語対応を推定する尺度としては，関連記事組における訳語候補の共起を利用する方法，あるいは，各言語のタームが出現する文脈の類似性を測定する等の有効性が示された [Utsuro03]． これまでの結果をふまえて，本稿では，それらの二種類 (実際には三種類) の尺度を統合する手法について述べる．具体的な統合方法としては，各尺度により訳語候補を順位付けした後，それらの順位を重み付きで足し合わせるという方法を用い，これにより，英語タームに対する日本語訳語推定の性能を改善できることを示す．

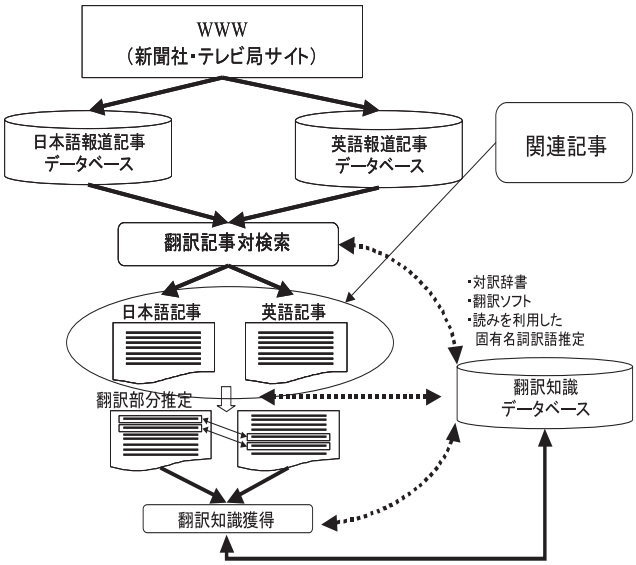

図 1: 日英関連報道記事からの翻訳知識獲得の流れ

2 言語横断関連報道記事検索

言語横断関連報道記事検索においては，まず，新聞社やテレビ局のサイトから英語記事 d_E と日本語記事 d_J を取得する．次に，関連記事対はお互いの日付が近いと想定して，日付の情報を用いて検索対象の記事を絞りこむ．そして，取得した英語記事 d_E と日本語記事 d_J の間の類似性を測るために，翻訳ソフト・対訳辞書などの情報源[1] を利用して英語記事 d_E を日本語訳に

*Combining Multiple Measures for Estimating Bilingual Term Correspondences from Japanese-English Relevant News Articles

[1] 5.4 節で述べるように，翻訳ソフト (オムロン社製「翻訳魂」) と対訳辞書 (英辞郎 Ver.37，85 万語) を比較した結果では，翻訳ソフトの方が高い検索性能を達成しており，そのため，訳語対応推定においても，翻訳ソフトを用いた方が高い性能が得られている．そ

変換し，この日本語訳と日本語記事 d_J から翻訳頻度ベクトル $v_{trJ}(d_E)$ と日本語頻度ベクトル $v(d_J)$ をそれぞれ作成する[2]．最後に，頻度ベクトル間で類似度を計算し，類似度が下限値以上の記事を検索結果とする．

ここで，この検索結果から，日英関連記事組を作成する場合には，英語記事を検索質問として関連日本語記事を収集する場合と，逆に，日本語記事を検索質問として関連英語記事を収集する場合の二通りが考えられる．英語記事を検索質問として関連日本語記事を収集する場合は，検索質問となる英語記事を d_E として，d_E の日本語訳頻度ベクトル $v_{trJ}(d_E)$ との間で余弦類似度の値が下限値 L_d 以上となる日本語記事の集合を D_J とする．

$$D_J \quad = \quad \left\{ d_J \mid \cos(v_{trJ}(d_E), v(d_J)) \geq L_d \right\}$$

そして，D_J 中の記事を結合することにより一つの日本語記事 D'_J を構成し，このような英日関連記事組 $\langle d_E, D'_J \rangle$ を集めた集合を RC_{EJ} とする．

$$RC_{EJ} \quad = \quad \left\{ \langle d_E, D'_J \rangle \mid D_J \neq \emptyset \right\}$$

同様に，日本語記事を検索質問として英日関連記事組を集めた集合を RC_{JE} とする．

3 日英関連報道記事における訳語対応の推定

本稿では，関連記事組の集合 RC_{EJ} もしくは RC_{JE} から訳語対応を推定する方法として，関連記事組の集合を疑似的な対訳コーパスとみなして，対訳コーパスにおける共起頻度を用いた訳語対応推定尺度を適用する方法，および，関連記事組の集合をコンパラブルコーパスとみなして，コンパラブルコーパスからの訳語対応推定手法を適用する方法の二種類を用いる．

以下，訳語対応推定の対象となる英語ターム (連語または単語) を t_E，日本語ターム (連語または単語) を t_J として，t_E と t_J の間の訳語対応推定値を $corr_{EJ}(t_E, t_J)$ とする．本稿では，t_E の品詞列としては任意のものを，また，t_J の品詞列としては,「茶筌」により品詞列を推定し，接頭詞，名詞，動詞によって構成される任意の列を対象としている．

3.1 関連記事組における訳語候補の共起および分割表を用いた推定

関連記事組の集合 RC_{EJ} もしくは RC_{JE} を疑似的な対訳コーパスとみなして訳語対応の推定を行う場合は，RC_{EJ}(または RC_{JE}) 中の関連記事組 $\langle d_E, D'_J \rangle$(または $\langle D'_E, d_J \rangle$) において t_E と t_J が共起する記事組数 $df(t_E, t_J)$，t_E のみが含まれ t_J が含まれない記事組数 $df(t_E, \neg t_J)$，t_J のみが含まれ t_E が含まれない記事組数 $df(\neg t_E, t_J)$，t_E も t_J も含まれない記事組数 $df(\neg t_E, \neg t_J)$ を用いて以下の 2×2 分割表を構成する．

	t_J	$\neg t_J$
t_E	$df(t_E, t_J) = a$	$df(t_E, \neg t_J) = b$
$\neg t_E$	$df(\neg t_E, t_J) = c$	$df(\neg t_E, \neg t_J) = d$

そして，以下の ϕ^2 統計を用いて t_E と t_J の統計的相関を測定し，訳語対応推定値 $corr_{EJ}(t_E, t_J)$ とする．

$$\phi^2(t_E, t_J) \quad = \quad \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

ここで，RC_{EJ} から求めた推定値を $\phi^2_{EJ}(t_E, t_J)$，RC_{JE} から求めた推定値を $\phi^2_{JE}(t_E, t_J)$ とする．

3.2 文脈の類似性を用いた推定

関連記事組の集合 RC_{EJ} もしくは RC_{JE} をコンパラブルコーパスとみなして訳語対応の推定を行う場合は，t_E および t_J についての文単位の文脈頻度ベクトルを求め，これらの文脈頻度ベクトル間の類似性を用いて t_E と t_J の訳語対応を推定する．具体的には，英語記事 d_E に対する日本語訳を作成しておき，d_E において t_E が出現する文の日本語訳の頻度ベクトルを加算して，t_E に対する文単位の文脈頻度ベクトル $cv_{trJ}(t_E)$ を構成する．同様に，日本語記事 d_J を集めた記事集合において t_J が出現する文について，それらの頻度ベクトルを加算することにより，t_J に対する文単位の文脈頻度ベクトル $cv(t_J)$ を構成する．そして，この文脈頻度ベクトル間の余弦 $\cos(cv_{trJ}(t_E), cv(t_J))$ を $corr_{EJ}(t_E, t_J)$ とする．この場合，RC_{EJ} もしくは RC_{JE} のどちらを用いても，訳語対応の推定結果は同じになる．

4 複数の訳語対応推定尺度の統合

英語ターム t_E に対して日本語訳語候補の順位付けを行うタスクにおいて，i) 英日方向の関連記事検索の結果から推定した ϕ^2 統計 ϕ^2_{EJ}，ii) 日英方向の関連記事検索の結果から推定した ϕ^2 統計 ϕ^2_{JE}，iii) 文脈ベクトルの類似性 $\cos(cv_{trJ}(\cdot), cv(\cdot))$，の三種類の尺度を統合する手法について述べる．具体的には，まず，英語ターム t_E に対して日本語訳語候補 t_J の順位付けを行い，各訳語対応推定尺度により推定した日本語訳語候補 t_J の順位を求める．

- 英日方向の ϕ^2 統計による順位

  $$r(t_J \mid t_E, \phi^2_{EJ})$$

  日本語訳語候補 t_J を，英語ターム t_E との間の訳語対応推定値 $\phi^2_{EJ}(t_E, t_J)$ の降順に整列したときの t_J の順位を用いる．

こで，本稿では，翻訳ソフトを用いて英語記事の日本語訳を行った後，関連記事検索を行った結果を用いる．

[2] 日本語形態素解析システム「茶筌」(http://chasen.aist-nara.ac.jp/) を用いて形態素列に分割し，平仮名語の高頻度機能的表現 26 語を不要語として削除した．また，単語頻度ベクトルは名詞と動詞のみを利用して生成した．なお，5 形態素以下の連語および単語の頻度を次元として頻度ベクトルを構成する場合と，単語の頻度のみを次元として頻度ベクトルを構成する場合を比較すると，翻訳ソフトを用いた場合では大きな差はない．しかし，対訳辞書を用いた場合は，連語および単語の頻度を次元とする方が性能がよいため，こちらを用いることとする．

表 1: 記事の日数・記事数・平均記事長

	総日数	総記事数	一日の平均記事数	一記事の平均記事長 (byte)
英語	935	23064	24.7	3228.9
日本語	941	96688	102.8	837.7

表 2: 記事間類似度の下限を満たす日英報道記事の数

類似度下限 L_d	CLIR なし	0.3	0.4	0.5
日付幅 (日)		± 2		
英語記事数	23064	6073	2392	701
日本語記事数	96688	12367	3444	882
日本語記事数 (重複あり)		16507	3840	918

- 日英方向の ϕ^2 統計による順位

$$r(t_J \mid t_E, -r(t_E \mid t_J, \phi^2_{JE}))$$

英語訳語候補 t_E を，日本語ターム t_J との間の訳語対応推定値 $\phi^2_{JE}(t_E, t_J)$ の降順に整列したときの t_E の順位を $r(t_E \mid t_J, \phi^2_{JE})$ として，この値の昇順に日本語訳語候補 t_J を整列したときの t_J の順位を用いる．

- 文脈ベクトルの類似性による順位

$$r(t_J \mid t_E, \cos(cv_{trJ}(\cdot), cv(\cdot))$$

日本語訳語候補 t_J を，英語ターム t_E との間の文脈ベクトル余弦 $\cos(cv_{trJ}(t_E), cv(t_J))$ の降順に整列したときの t_J の順位を用いる．

これらの三種類の尺度による順位を統合した結果の順位 $r(t_J \mid t_E, r_{comb})$ としては，以下の重み付き和 $r_{comb}(t_J \mid t_E)$ の昇順に整列したときの t_J の順位を用いる．ただし，重み α_1，α_2 については，訓練用データより最適値を求める．

$$\begin{aligned} r_{comb}(t_J \mid t_E) \quad \equiv \quad & \alpha_1 \cdot r(t_J \mid t_E, \phi^2_{EJ}) \\ & +\alpha_2 \cdot r(t_J \mid t_E, -r(t_E \mid t_J, \phi^2_{JE})) \\ & +(1-\alpha_1-\alpha_2) \cdot r(t_J \mid t_E, \cos(cv_{trJ}(\cdot), cv(\cdot))) \end{aligned}$$

5 実験および評価

5.1 報道記事セット

国内の新聞社のサイト (今回の実験では一つのサイト) から，表 1 に示す日数・記事数・記事長の日本語および英語の報道記事を収集した．次に，[堀内 02] における調査結果に基づいて，英語の記事に対してほぼ同一の内容の日本語記事が存在する日付の幅を設定し，その日付の幅の範囲で言語横断関連報道記事検索を行った．記事間類似度下限 L_d を変化させた場合に検索される記事数の一覧を表 2 に示す．ここで,「日本語記事数 (重複あり)」の欄には，二つ以上の英語記事に対して重複して検索された日本語記事を重複して数えた記事数を示す．この結果から，類似度下限 L_d が 0.4 や 0.5 の場合は，利用可能な記事数が著しく減少することが分かる．予備実験においては，訳語対応推定が安定して行えるためには，一定規模以上の記事が必要であるという結果が得られていたため，以降の訳語対応推定は，類似度下限 $L_d = 0.3$ の条件のもとで行う．

5.2 評価用英語タームの選定

今回の評価実験では，評価用英語ターム (単語または連語) を人手で選定しておき，評価用英語タームに対する日本語訳語候補の順位付けの性能の評価を行った．実装の都合上，英語タームおよび日本語タームを構成する単語数に上限 U^E_l および U^J_l を設け，$U^E_l = U^J_l = 5$ とした．評価用英語タームを選定するために，まず，英語タームとしての適切さの尺度として C-value [Frantzi00] を用い，C-value の上位 4,000 個の単語列を抽出した．ここでは，C-value の計算に用いる頻度の計算を効率よく行うために，PrefixSpan [Pei01][3] を用いて頻度 10 以上の単語列の頻度を測定した．次に，単語列の上で包含関係にある単語列同士をグルーピングし，2,753 個の英語タームグループを作成した．そして，ある英語タームのグループについて，その要素となる英語ターム t_E が任意の日本語訳語候補に対して持つ英日方向 ϕ^2 統計値 $\phi^2_{EJ}(t_E, t_J)$ の最大値を，そのグループの持つ ϕ^2 統計値とみなして，2,753 個の英語タームグループを ϕ^2 統計値の降順に整列した．この整列済み英語タームグループの上位 1,000 グループから，i) グループの上位から順に評価用英語タームを 100 個選定したもの，ii) 上位 1,000 グループから，無作為に評価用英語タームを 100 個選定したもの，という二つの評価用英語タームセットを作成した[4]．評価実験の結果では，下位のグループから選定した英語タームほど，訳語候補順位付けの性能が下がる傾向となったが，二つのタームセットの間では，性能の差は数%程度であった．そこで，以下では，セット i) を用いた実験結果のみ示す．

5.3 訳語対応推定の性能

4 節で述べた三種類の尺度それぞれを比較すると，英日方向の ϕ^2 統計 ϕ^2_{EJ} が最も高精度でかつ計算コストも小さい．そこで，まず，英日方向の ϕ^2 統計 ϕ^2_{EJ} により順位付けした上位 30 個の日本語訳語候補[5] を求め，

[3] http://cl.aist-nara.ac.jp/~taku-ku/software/prefixspan/

[4] いずれも，本稿で用いた翻訳ソフトおよび対訳辞書では訳せないタームから構成される．

[5] さらに，ϕ^2 統計値が同一で包含関係にある日本語タームの組に対して，包含されるタームを削除するというフィルターをかける．

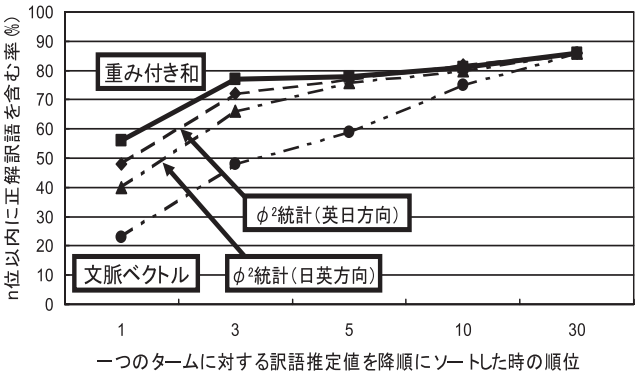

図 2: 複数の訳語対応推定尺度および重み付き和の比較

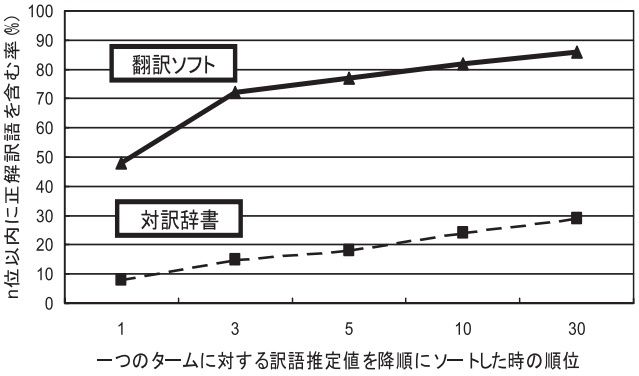

図 3: 翻訳ソフトと対訳辞書の比較

この訳語候補に対して，他の二つの尺度による順位を求めた．そして，4 節の統合法によりこれらの三つの順位の統合を行った．統合の際には，全英語タームの 4/5 で重み α_1，α_2 を決定し，残りの 1/5 で精度を評価することとし，5 分割交差検定によって精度を測定した．上位 n 位以内に正解訳語 (今回の実験では，各英語タームにつき一つだけ) が含まれる英語タームの割合をプロットした結果を図 2 に示す．

単独の尺度としては，英日方向の ϕ^2 統計の性能が最も高く，日英方向の ϕ^2 統計，文脈ベクトルの順に性能が低くなる．ϕ^2 統計と比較すると，文脈ベクトルの性能はかなり低くなる．これらの三つの尺度を統合した結果では，特に順位の上位の部分において，英日方向の ϕ^2 統計の性能を改善することができた．また，このときの各尺度の重みは，(英日ϕ^2 : 日英ϕ^2 : 文脈ベクトル) = (0.6 : 0.4 : 0) となり，文脈ベクトルによる訳語対応推定結果は考慮しない方がよいという結果となった．ただし，セット ii) の場合には，各尺度の重みは，(英日ϕ^2 : 日英ϕ^2 : 文脈ベクトル) = (0.5 : 0.4 : 0.1) となり，三つの尺度全てを考慮した統合が最適となった．

5.4 英語記事の翻訳方法の比較

訳語対応推定尺度として英日方向の ϕ^2 統計を用いた場合について，2 節の日英関連報道記事検索において，英語記事の翻訳に翻訳ソフトを用いた場合と対訳辞書を用いた場合の比較を行った結果を図 3 に示す．言語横断関連報道記事検索においては，対訳辞書を参照して可能な訳を全て列挙するのではなく，翻訳ソフトを用いて一意の訳語選択を行った方が，関連報道記事検索の精度は高くなる．このため，翻訳ソフトの方が圧倒的に高い性能となったと考えられる．

6 おわりに

本稿では，日英関連報道記事からの訳語対応推定のタスクにおいて，訳語対応を推定する複数の尺度を統合する手法について述べ，実際に，英語タームに対する日本語訳語推定の性能が改善できることを示した．今後は，ウェブ検索エンジンにより収集した日英非対訳文書より得られる訳語候補の順位付け [木田 04] と，日英報道記事から得られる順位付けの統合を行い，訳語対応推定の性能の改善を試みる予定である．

謝辞: PrefixSpan の適用において協力頂いた奈良先端大 工藤拓氏に感謝する．

参考文献

[Frantzi00] Frantzi, K., Ananiadou, S. and Mima, H.: Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method, *Inter. J. Digital Libraries*, Vol. 3, No. 2, pp. 115–130 (2000).

[Fung98] Fung, P. and Yee, L. Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts, *Proc. 17th COLING and 36th ACL*, pp. 414–420 (1998).

[浜本 03] 浜本武, 中山健明, 日野浩平, 堀内貴司, 宇津呂武仁：言語横断関連報道記事検索における翻訳ソフト・対訳辞書・数値表現翻訳規則の性能比較, 言語処理学会第 9 回年次大会論文集, pp. 425–428 (2003).

[堀内 02] 堀内貴司, 千葉靖伸, 浜本武, 宇津呂武仁：言語横断検索により自動収集された日英関連報道記事からの訳語対応の獲得, 情報処理学会研究報告，2002–NL–150, pp. 191–198 (2002).

[木田 04] 木田充洋, 宇津呂武仁, 日野浩平, 佐藤理史：ウェブ上の日英非対訳文書を用いた訳語対応推定, 言語処理学会第 10 回年次大会論文集 (2004).

[Matsumoto00] Matsumoto, Y. and Utsuro, T.: Lexical Knowledge Acquisition, Dale, R., Moisl, H. and Somers, H. (eds.), *Handbook of Natural Language Processing*, chapter 24, pp. 563–610, Marcel Dekker Inc. (2000).

[Pei01] Pei, J., Han, J., Mortazavi-Asl, B. and Pinto, H.: PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth, *Proc. Inter. Conf. Data Mining*, pp. 215–224 (2001).

[Utsuro03] Utsuro, T., Horiuchi, T., Hamamoto, T., Hino, K. and Nakayama, T.: Effect of Cross-Language IR in Bilingual Lexicon Acquisition from Comparable Corpora, *Proc. 10th EACL*, pp. 355–362 (2003).